Natural Language Engineering **23** (3): 441–472. © Cambridge University Press 2016
doi:10.1017/S1351324916000097

*NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic**

MAI OUDAH[1] and KHALED SHAALAN[2]

[1]*Masdar Institute of Science and Technology, Abu Dhabi, UAE*
e-mail: moudah@masdar.ac.ae
[2]*The British University in Dubai, Dubai International Academic City, Dubai, UAE*
e-mail: khaled.shaalan@buid.ac.ae

(*Received 13 September 2014; revised 24 March 2016; accepted 29 March 2016; first published online 6 May 2016*)

Abstract

Named Entity Recognition (NER) is an essential task for many natural language processing systems, which makes use of various linguistic resources. NER becomes more complicated when the language in use is morphologically rich and structurally complex, such as Arabic. This language has a set of characteristics that makes it particularly challenging to handle. In a previous work, we have proposed an Arabic NER system that follows the hybrid approach, i.e. integrates both rule-based and machine learning-based NER approaches. Our hybrid NER system is the state-of-the-art in Arabic NER according to its performance on standard evaluation datasets. In this article, we discuss a novel methodology for overcoming the coverage drawback of rule-based NER systems in order to improve their performance and allow for automated rule update. The presented mechanism utilizes the recognition decisions made by the hybrid NER system in order to identify the weaknesses of the rule-based component and derive new linguistic rules aiming at enhancing the rule base, which will help in achieving more reliable and accurate results. We used ACE 2004 Newswire standard dataset as a resource for extracting and analyzing new linguistic rules for person, location and organization names recognition. We formulate each new rule based on two distinctive feature groups, i.e. Gazetteers of each type of named entities and Part-of-Speech tags, in particular noun and proper noun. Fourteen new patterns are derived, formulated as grammar rules, and evaluated in terms of coverage. The conducted experiments exploit a POS tagged version of the ACE 2004 NW dataset. The empirical results show that the performance of the enhanced rule-based system, i.e. NERA 2.0, improves the coverage of the previously misclassified person, location and organization named entities types by 69.93 per cent, 57.09 per cent and 54.28 per cent, respectively.

1 Introduction

Named Entity Recognition (NER) is the task of identifying proper names in a given text and classifying them into predefined types, such as person, location and organization names (Nadeau and Sekine 2007; Shaalan 2014). In the 1990s, at the Message Understanding Conferences (MUC), the task of NER was firstly introduced and given attention by the community of research. Three main NER subtasks were

* This research was partially funded by the British University in Dubai (Grant No. INF004-Using machine learning to improve Arabic named entity recognition).

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

defined at the sixth MUC: ENAMEX (person, location and organization), TIMEX (temporal expressions) and NUMEX (numerical expressions). A customized NER system may consider more sub-types in the NER subtasks to fulfill the system objectives, e.g. Location NEs may have sub-types such as city, country, river, road, etc. NER is considered a critical pre-processing step in many Natural Language processing (NLP) applications. Examples of such NLP applications (Cowie and Wilks 1996) are Information Retrieval (Darwish and Magdy 2014), Machine Translation (Babych and Hartley 2003), Question Answering (Mesfar 2007; Hamadene, Shaheen and Badawy 2011; Abouenour, Bouzoubaa and Rosso 2012) and Text Clustering (Benajiba, Diab and Rosso 2009a). Most of NER systems are developed using either rule-based or machine learning (ML)-based approaches (Shaalan 2014). A third kind of approaches is called the hybrid approach, which integrates both rule-based and ML-based approach in order to overcome their weaknesses that arise because of the knowledge elicitation bottleneck and the lack of resources when it comes to some languages, such as Arabic.

Arabic language, an important member of the family of Semitic languages, is the official language in the Arab world countries where more than 300 million people speak Arabic in their daily life (Farghaly and Shaalan 2009; Shaalan 2010). Arabic is one of the richest natural languages in the world in terms of morphology (Al-Sughaiyer and Al-Kharashi 2004; Shaalan, Monem and Rafea 2007). Interest in Arabic NLP has been gaining momentum during the last decade, especially due to the challenges posed by the language's scarce resources (Shaalan and Raza 2009) and critical aspects, such as the complex morphology, the absence of capitalization, the common words versus proper noun (PN) ambiguities and the variant writing styles (Abdul-Hamid and Darwish 2010; Habash 2010; Shaalan 2014).

Recently, NER for Arabic has been receiving increasing attention, yet opportunities for improvement in performance are still available. Most of the work in the field of Arabic NER have been adopting either (1) the rule-based approach, notably NERA system (Shaalan and Raza 2008), or (2) the ML-based approach, notably ANERsys 2.0 (Benajiba and Rosso 2007). Each approach has its strengths and weaknesses. The rule-based approach uses handcrafted grammatical rules acquired from the linguists. For example, Azmi and Badia (2010) have proposed the use of domain specific grammar to parse narrator names in Prophet Mohammed's hadith text. Therefore, any maintenance applied to a rule-based system is labor-intensive and time-consuming, especially if a linguist with the required knowledge and background is unavailable. On the contrary, the ML-based approach utilizes statistical learning models that make use of selected features extracted from annotated training datasets. The main advantage of the ML-based systems over the rule-based systems is that they are adaptable and updatable with minimal time and effort as long as sufficiently large datasets are available. In the literature, the hybrid approach is adopted for NER to overcome the challenges imposed by the two other approaches when utilized separately, and as a result achieving better performance (Shaalan 2014).

Combining the rule-based approach with the ML-based approach for Arabic NER has resulted in the construction of the hybrid NER pipelined approach (Abdallah, Shaalan and Shoaib 2012; Oudah and Shaalan 2012; Oudah and Shaalan 2013). Our

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

state-of-the-art Arabic hybrid NER system (Shaalan and Oudah 2014) is capable of recognizing eleven different NE types, i.e. person, location, organization, date, time, price, measurement, per cent, phone number, ISBN and filename. The proposed hybrid approach, which is domain independent, has been applied to a number of Arabic standard datasets to evaluate its overall performance (Shaalan and Oudah 2014). Through pipelining, the Arabic NER task is divided into two stages that are processed sequentially: (1) rule-based NER processing that produces NE labels based on pre-defined lists of NEs and grammar rules, and (2) an ML post-processing that makes use of language-specific and language-independent features. The feature set utilized by the ML-based component includes the rule-based features, i.e. the output of the rule-based NER processing and the main contribution of the rule-based component to the hybrid approach formation. According to the experimental results, the hybrid NER system for Arabic outperforms the other state-of-the-art Arabic NER systems when applied to the standard dataset *ANERcorp* (Benajiba and Rosso 2007; Shaalan and Oudah 2014).

In this study, the main objective is to enhance the coverage and performance of the rule-based NER system. The first version of the rule-based component, i.e. the first stage in our Arabic hybrid NER system, is a reproduction of NERA. NERA is an Arabic rule-based NER system developed by Shaalan and Raza (2008; 2009). Two types of linguistic resources are collected and acquired, including gazetteers, i.e. pre-defined lists of NEs or keywords, also known as dictionaries, and corpora, i.e. training and testing datasets. The output of the rule-based component is then fed to the ML-based component after being transformed into a set of features, so-called rule-based features, along with the rest of considered features. Improving the rule-based system will allow for more reliable and accurate results when either the rule-based approach is used alone or as part of the hybrid approach. Moreover, an automated mechanism for rule-based system's maintenance will help eliminating the main disadvantage of the rule-based approach, i.e. being labor-intensive and time-consuming when it comes to updating the system.

In this article, we present a novel methodology for improving the rule-based component (NERA), which produces NERA 2.0, i.e. an enhanced version of NERA. The first step is to capture the cases where the NERA's grammar is weak, i.e. the cases NERA misclassifies but the hybrid system correctly classifies. The second step is to build new linguistic rules via analyzing the feature vectors of those cases. The final step is to implement the newly produced linguistic rules and then test their capabilities. We target the most distinctive and informative features of NEs from the set of features in our system, i.e. Gazetteers and Part-of-Speech (POS) tags (Shaalan and Oudah 2014). In this study, we identify fourteen new patterns that are produced by the proposed methodology. These patterns are distributed among four categories based on two factors: (1) the POS tag of the candidate word, whether it is a PN or a common noun (N), and (2) its presence status in the system's gazetteers. We conducted a number of experiments to evaluate the coverage of NERA 2.0. The dataset used for evaluation is a POS-tagged version of the Automatic Content Extraction (ACE) 2004 Newswire (NW), i.e. the NW data from the Arabic training dataset for the ACE evaluation conducted in 2004, which has been created by

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Linguistic Data Consortium (LDC) and is composed of 63,122 words with 6,627 ENAMEX NEs (Mitchell *et al.* 2005). ACE 2004 NW dataset is significant for two reasons. First, it is the largest dataset among ACE linguistic datasets for Arabic NER that has been widely used in the literature for training and testing Arabic NER systems in terms of the number of present named entities (NEs) (i.e. 6,627 of ENAMEX NEs). Second, it is a standard linguistic resource used by many remarkable Arabic NER systems for evaluation. According to the experimental results, NERA 2.0 is capable of obtaining an increased coverage of 69.93 per cent, 57.09 per cent and 54.28 per cent for person, location and organization names, respectively, of those previously misclassified by NERA.

The structure of the article is as follows. Section 2 gives a brief background about Arabic language characteristics. Literature review of NER is given in Section 3. Section 4 illustrates the architecture of our hybrid NER framework. Section 5 describes the methodology adopted for acquiring and applying new rules. The experiments are discussed in Section 6. Section 7 concludes this article and gives directions for future research.

2 Arabic language characteristics

Based on the language usage, Arabic can be classified into three types: Classical Arabic, Modern Standard Arabic and Colloquial Arabic Dialects. Classical Arabic is the formal version of Arabic, which has been used for centuries as the language of Islam, and Muslims, including Non-Arabs, use it in their daily prayers. Modern Standard Arabic is the language of today's Arabic media and education. The main difference between Modern Standard Arabic and Classical Arabic lies in the introduction of modern vocabulary, the needs of contemporary expression and the exclusion of short vowels (Zaghouani 2012). Colloquial Arabic is the spoken Arabic used by Arabs in their informal daily communication and are not taught in schools due to its irregularity. There is a great variety in this branch of Arabic, as every country in the Arab world has almost its own dialect, which can be incomprehensible to a speaker of another dialect. In Salloum and Habash (2012), the authors presented a universal machine translation pre-processing approach that has the ability to produce Modern Standard Arabic (MSA) paraphrase of dialectal input. In this way, available MSA tools can be used also to process Colloquial Arabic text as most of the Arabic NER systems are developed to support MSA.

Applying NLP tasks, in general, and the NER task, in particular, is very challenging when it comes to Arabic because of its peculiarities and unique nature. The main characteristics of Arabic that pose non-trivial challenges for NER task include:

- *No capitalization:* Capitalization is not a feature of Arabic script; unlike Latin languages where an NE usually begins with a capital letter, e.g. Egypt versus مصر (Misr). Therefore, the usage of the capitalization feature is not an option in Arabic NER. However, when the English (gloss) translation of the Arabic

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

word is available, it can be exploited to indicate the presence of an NE (Farber *et al.* 2008).

- *The agglutinative nature:* Arabic language has a high agglutinative nature in which a word may consist of prefixes, lemma and suffixes in different combination, which results in a very complicated morphology (AbdelRahman *et al.* 2010). For example, the word وبقصصهم *wabiqiSaSihim* 'and by their stories', according to the Habash–Soudi–Buckwalter transliteration scheme (Habash, Soudi and Buckwalter 2007), consists of *wa* 'and' as a conjunction, *bi* 'by' as a preposition (i.e. prefixes), *qiSaSi* 'stories' as the stem, and *him* 'their' as a possessive pronoun (i.e. suffix). Morphological analysis can be utilized for tackling this challenging characteristic.
- *Optional short vowels:* In theory, short vowels, or diacritics, are needed for pronunciation and disambiguation (Shaalan *et al.* 2009; Habash and Rambow 2005). However, practically, most Modern Standard Arabic texts do not include diacritics, and hence, the surface form of a word in Arabic may refer to different entities according to the context they appear in, which results in creating one-to-many ambiguity. For instance, قطر may refer to the country name 'Qatar' (a location NE) if transliterated as 'Qatar', the literal meaning of country (a trigger word for location NEs) or 'radius' (a trigger word for measure NEs) if transliterated as 'qutr', or the literal meaning of distill if transliterated as 'qat∼ar'. Unfortunately, this solution may not work if the contextual information is itself ambiguous due to non-vocalization (Mesfar 2007). This challenge can be dealt with via morphological and contextual analysis of the candidate words and their context.
- *Spelling variants:* In Arabic script, a word can be spelled differently and still refers to the same orthographic form with same meaning, creating a many-to-one ambiguity (Shaalan and Raza 2009). For example, the person name (Rola) can be written as رولى or رولا, and the capital name (Abu Dhabi) that can be written as أبوظبي or ابوظبي. Morphological and contextual analysis can help overcome this challenge too, especially when the POS tag is utilized.
- *Lack of linguistic resources:* In general, there is a limitation in the number of Arabic linguistic resources that are publicly available free for the research purposes. In particular, many of the available corpora are neither annotated with NEs nor have sufficient number of NEs, which make them unsuitable for training and/or evaluating Arabic NER systems (Shaalan 2014). Therefore, researchers usually have to spend tangible efforts to acquire/annotate and verify their own Arabic linguistic resources in order to properly develop Arabic NER systems (Benajiba and Rosso 2007).

3 Literature review of NER

NER serves to achieve two main goals, including the detection of NEs and the classification of those NEs into a set of predefined types. In the literature, to fulfill those two goals, three types of approaches are used: rule-based, ML-based

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

and hybrid approaches. They are presented in the following subsections in their chronological order.

3.1 Rule-based NER

Rule-based NER systems depend on local handcrafted linguistic rules to identify NEs within the text using linguistic and contextual clues, and indicators (Shaalan 2010). Such systems exploit the gazetteers/dictionaries as auxiliary clues to the rules. The rules are usually implemented in the form of regular expressions or finite-state transducers (Mesfar 2007). The maintenance of a rule-based system is not a straightforward process as experienced linguists need to be available to provide the system with the proper adjustments (Mayfield, McNamee and Piatko 2003). Thus, any serious update to this kind of systems is labor-intensive and time-consuming.

Maloney and Niv (1998) have presented TAGARAB system, which is one of the early attempts to tackle Arabic NER. It is a rule-based system where a pattern matching engine is combined with a morphological tokenizer to recognize person, organization, location, number and time NEs. The empirical results show that combining NE recognizer with a morphological tokenizer outperforms the individual NE recognizer in terms of accuracy when applied to random datasets from AI-Hayat newspaper.

Maynard *et al*. (2001) have introduced MUSE system, which is a grammar-based NER system implemented in the GATE framework. MUSE is composed of a tokenizer, gazetteers and grammars. After a preliminary analysis of the input text, the set of processing resources is manually manipulated in order to extract NEs based on the text type. The evaluation was conducted on English datasets to extract NEs of various types including person, date, location, organization, time, money, per cent, email, URL, telephone, IP and identifier.

Another work adopting the rule-based approach for NER is the one developed by Shaalan and Raza (2007) called PERA. PERA is a grammar-based system which is built for identifying person names in the Arabic script with high degree of accuracy. PERA is composed of three components: gazetteers, local grammar and a filtration mechanism. Whitelists of complete person names are provided in the gazetteer component in order to extract the matching names regardless of the grammars. Afterwards, the input text is presented to the local grammar, which is in the form of regular expressions, to identify the rest of Person NEs using gazetteers and indicators. Finally, the filtration mechanism is applied on the NEs detected through certain grammatical rules in order to exclude the ambiguous and invalid NEs. PERA achieved satisfactory results when applied to the ACE and the Treebank Arabic datasets.

Mesfar (2007) has developed an Arabic component under the NooJ framework, i.e. a freeware linguistic environment that is based on .NET platform and has an object-oriented architecture that enables the Arabic text processing and NER. The NER component consists of a tokenizer, a morphological analyzer and an NE finder. The NE finder exploits a set of gazetteers and indicator lists to support the rule construction. The system identifies NEs of the types: person,

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

location, organization, currency and temporal expressions. The system utilizes the morphological information to extract unclassified PN and thereby enhance the overall performance of the system.

As a continuation of Shaalan and Raza (2007) research work, NERA system was introduced (Shaalan and Raza 2008; Shaalan and Raza 2009). NERA is a rule-based system that is capable of recognizing NEs of ten different types: person, location, organization, date, time, ISBN, price, measurement, phone numbers and filenames. The implementation of the system was in the FAST ESP framework, where the system has the same three components as in the PERA system with the same functionalities to cover the ten NE types. The authors have constructed their own corpora from different resources in order to have a representative number of instances for each NE type. According to their experimental results, the system achieves satisfactory accuracy scores when applied to the developed corpora. As part of our hybrid system, we have constructed a rule-based component, which implements the NERA grammatical rules along with additional rules to cover more types of NEs.

Elsebai, Meziane and BelKredim (2009) have proposed a rule-based NER system that integrates pattern matching with morphological analysis to extract person names from Arabic text. The pattern matching engine utilizes lists of keywords without using pre-defined lists of Person names. The system has achieved an F-measure of eighty-nine per cent when applied to 700 news articles obtained from Aljazeera television website.

Riaz (2010) has introduced a rule-based system, which tackles the task of Urdu NER. Urdu is a language written in a script that is derived from Arabic. Due to the lack of large annotated datasets, linguistic rules were constructed instead of training a statistical model to extract NEs of types: person, location, organization, date and number. The system outperforms the Conditional Random Fields (CRF)-based NER systems proposed in IJCNLP 2008 as reported by Riaz (2010).

Zaghouani (2012) has introduced a rule-based Arabic NER system (RENAR) to extract person, location and organization NEs. The system is composed of three phases: (1) morphological pre-processing, (2) looking up known NEs and (3) using local grammar to extract unknown NEs. According to the empirical results, RENAR outperforms ANERsys 1.0 (Benajiba, Rosso and Bened'I 2007), ANERsys 2.0 (Benajiba and Rosso 2007) and LingPipe (Alias-I 2008) in extracting Location NEs when applied to ANERcorp dataset, while LingPipe outperforms RENAR in extracting person and organization NEs.

Aboaoga and Aziz (2013) have introduced a rule-based Arabic NER system that extracts person names from Arabic text. The domains covered by the system are sports, politics and economics. The authors have developed four main linguistic rules, which compose the core of the system. The system goes through three processing steps for Arabic person names recognition including (1) pre-processing (tokenization, data cleaning and sentence splitting), (2) automatic NE tagging where pre-defined lists of person names and keywords are used for person NE annotation process and keyword annotation process and (3) applying the rules to the text in order to extract person names that do not exist in the built-in dictionaries. The dataset used

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

for evaluation is collected from different online Arabic newspapers. The empirical results showed that the system achieved accuracy in the sport domain higher than the ones achieved in the politic and economic domains.

3.2 Machine learning-based NER

ML-based NER systems utilize ML algorithms in order to learn NE patterns from annotated texts. The most common approach used for ML-based NER is the supervised learning (SL) approach, which handles the NER problem as a classification problem and requires the availability of large and sufficient annotated datasets. Among the most common supervised learning algorithms utilized for NER are Support Vector Machines (SVM), CRF, Maximum Entropy (ME), Hidden Markov Models and Decision Trees (Collins 2002; Nadeau and Sekine 2007).

Zhou and Su (2002) have adopted Hidden Markov Model as the ML approach in their NER system, which has the ability of recognizing different types of NEs including proper names, time and numerical expressions in English text. The utilized feature set includes features of two categories including internal evidence and external evidence. The internal evidence focuses on word-level features e.g. InitialCaps, Semantic features e.g. SuffixPercent (%), and Gazetteer features that determine the presence of the candidate word in the system's gazetteers. On the other hand, the features representing the external evidence are the contextual features derived from the context to extract the relationships between previously recognized entities and the candidate words in the input text. The system has been applied on MUC-6 and MUC-7 English shared tasks and the results show that the performance reaches its highest rates when all the features from both categories are considered by the system.

Mayfield *et al.* (2003) have proposed a language-independent statistical-based system for NER. The system exploits SVM approach to handle large amount of features with minimal overfitting. Examples of the features are word's length, POS tag, whether the word is between double quote marks or not, and whether a dash exists in the word or not. The features of the surrounding words contribute to the learning and classification processes as well. The window size is seven words centered by the candidate word. The system has been applied on CoNLL2003 English and German datasets for training and evaluation.

Benajiba *et al.* (2007) have developed an Arabic NER system, ANERsys 1.0, which uses ME. The authors have built their own linguistic resources, which have become a *de facto* standard in Arabic NER literature, i.e. ANERcorp (an annotated corpus) and ANERgazet (person, location and organization gazetteers). The features used by the system are lexical, contextual and gazetteers features. The system can recognize four types of NEs: person, location, organization and miscellaneous. The ANERsys 1.0 system faced some difficulties when detecting NEs of multiple tokens/words and hence, Benajiba and Rosso (2007) have developed ANERsys 2.0, which employs a two-step mechanism for NER: (1) detecting the boundaries of NEs, and (2) identifying the NE types. Benajiba and Rosso (2008) have applied CRF instead of ME in attempt to improve the performance. Both of the ANERsys

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

2.0 and the CRF-based system share the same set of features to exploit for ML, including: POS tags, base phrase chunks, gazetteers, and nationality features. The CRF-based system achieves higher results in terms of accuracy.

Benajiba, Diab and Rosso (2008a) investigated the sensitivity of different NE types to various types of features. They build multiple classifiers for each NE type adopting SVM and CRF approaches. ACE datasets are used in the evaluation process. According to their findings, it cannot be stated whether CRF is better than SVM or vice versa in Arabic NER. Each NE type is sensitive to different features and each feature plays a role in recognition process. Further studies, i.e. Benajiba, Diab and Rosso (2009a, 2009b), have confirmed the importance of considering language-independent and language-specific features in Arabic NER as well.

Benajiba, Diab and Rosso (2008b) have developed another NER system based on SVM. The features used are contextual, lexical, morphological, gazetteers, POS tags and BPC, nationality and the corresponding English capitalization. The system has been evaluated using ACE Corpora and ANERcorp. The best results are achieved when all the features are considered by the system.

Finkel and Manning (2009) have introduced a new approach to enhance the recognition of nested NEs. NEs can be nested if one NE or more are within the structure of other NEs. For example, 'The British University in Dubai', shows location name, Dubai, embedded in the organization name. The proposed approach relies on representing the sentences within a text using constituency parse trees where POS tags are included, and then a discriminative constituency parser (i.e. CRF-based) is used to generate the classification models. Each node within the tree needs to be labeled with its parent and grandparent. The system, which is domain-independent, was applied on GENIA v.3.02, JNLPBA and AnCora datasets. The empirical results show that the overall performance of the new approach is better than the one of semi-CRF model in terms of the standard evaluation metrics when used to extract NEs of any level, i.e. either nested or normal NEs.

Mohammed and Omar (2012) have proposed an ML-based system, which adopts Artificial Neural Networks approach, to tackle the problem of Arabic NER. The system has three main phases including (1) Pre-processing of the data (data cleaning, text tokenization and tagging using Inside-Outside-Beginning schema), (2) Romanization of Arabic, i.e. rendering written Arabic in the Latin script and (3) applying the Artificial Neural Networks classifier to the text. The system is capable of recognizing four NE types including person, location, organization and miscellaneous. ANERcorp dataset is used for evaluation. The authors have compared the performance of the system when decision trees and Artificial Neural Networks approaches are used individually. The relationship between the corpus size and the accuracy of the system has been examined as well. According to the experimental results, when Artificial Neural Networks approach is adopted, it achieved higher accuracy than the decision trees approach. The results showed that the accuracy improves as the size of the corpus increases.

Zirikly and Diab (2015) have proposed an NER system for Arabic that can handle text from the web; social media, in particular. They adopt supervised learning, represented by CRF algorithm, to learn a classifier for labeling Dialectal Arabic text.

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Their feature set includes Lexical Features, Contextual Features and Morphological Features. According to their experimental analysis, the proposed system outperforms state-of-the-art results for Dialectal Arabic, when applied to Dialectal Arabic dataset, i.e. an annotated data from web blogs identified by LDC as Egyptian dialect, with F1-score of 72.68 per cent.

3.3 Hybrid NER

The hybrid approach is the integration between the rule-based approach and the ML-based approach (Petasis *et al.* 2001). The direction of the processing flow can be from the rule-based system to the ML-based system or vice versa. In the literature, the hybrid approach has been utilized in NER for different languages, such as English, French and Chinese.

Srihari, Niu and Li (2000) have proposed a hybrid NER system, which exploits two ML algorithms and pattern matching rules to extract NEs of several types, including person, location, organization, date, money, time and percentage from English text. The system is composed of four main modules. The first module is the pattern matching rules which is used to extract temporal and numerical expressions. The second module is a ME model that is built to utilize the gazetteers and the contextual information in order to provide a preliminary recognition of person and location NEs. The third module is the Hidden Markov Model classifier which gives the final tagging for the entities of types: person, location and organization. The final module is another ME model produced to identify further sub-categories such as Government and Airport. The system is evaluated using MUC-7 dataset and the results show that using the hybrid approach improves the performance in terms of accuracy. Our hybrid approach is mainly different from this approach in two aspects. First, the number of modules within their architecture requires more processing. Second, the modules in their system do not share the capability of handling the same types of NEs.

Seon *et al.* (2001) have introduced a hybrid approach to tackle the problem of NER for Korean. Two ML approaches are used by the system: ME and Neural Networks. The ME is utilized to resolve the problem of unknown words that do not exist in any of the pre-defined dictionaries within the system, while Neural Networks use lexical information to deal with ambiguity. The pattern-selection rules are exploited to just combine adjacent words that comprise a multiword NE. The system is developed to extract NEs of types: person, location and organization. In comparison, the authors have utilized pure ML learning approach for the actual NER while the role of the rule-based post-processing is not NER but rather setting the boundary of the identified NEs by combining them which might be suitable for Korean NER.

Petasis *et al.* (2001) have introduced a hybrid method for the purpose of facilitating the maintenance of their rule-based NER system. The role of the ML-based component is to evaluate the performance of the rule-based NER system, and accordingly it takes the decision of whether to maintain the rule-based NER system. In the training phase, the ML-based system relies only on the annotated output

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

of the rule-based system and hence, there is no need to train on manually tagged dataset. Then, the two systems are applied on a new dataset and their results are compared. The disagreements in recognition output indicate the need to update the rule-based system. The ML technique adopted is C4.5 that implements decision tree ML method. Greek and French rule-based systems are evaluated using this approach. The systems can recognize NEs of types: person, location and organization. The experimental analysis shows that this method can help in deciding when to maintain a rule-based NER system. Our hybrid approach is mainly different from this approach in two aspects. First, the authors' approach is mainly rule-based approach which employs ML approach for the purpose of enhancing the performance of the rule-based approach while our architecture is hybrid and does not favor one approach over the other. Second, the authors' approach restricts the feature extraction to the rule-based output (annotations) to avoid the need of large annotated corpus for ML.

Mencius is a hybrid NER system for Chinese proposed by Tsai *et al.* (2004). The rule-based and ML-based approaches are combined in Mencius in order to enhance the recognition capabilities for person, location and organization NEs. The output of the rule-based module is used as an input to the ML-based module, which is a ME Model. The feature set consists of internal (i.e. category-dependent) features utilized to distinguish between different NE categories. The authors have built their own corpus to evaluate the performance of their system. According to the empirical results, the highest performance is achieved when the hybrid system is used along with word tokenization.

Küçük and Yazici (2012) have presented a hybrid NER system for Turkish. The main module in the system is the rule-based module which includes knowledge sources for specific domains. The original rule-based system is extended to a hybrid NER system which learns from annotated text with the purpose of improving the knowledge sources accordingly. The authors have developed their own corpus for learning and evaluation purposes. The rule-based module contains dictionaries of person, location and organization NEs along with linguistic patterns to recognize the previously mentioned NE types as well as the numerical and temporal expressions. The system utilizes rote learning for ML (Küçük and Yazici 2012). According to the empirical results, the hybrid NER system for Turkish outperforms the pure rule-based NER system in terms of accuracy. Our proposed hybrid approach differs from this approach in several aspects. The flow of processing in Küçük and Yazici's approach goes from the ML-based module to the rule-based module, while it is vice versa in our approach. In the authors' system, the rule-based module is heavily dependent on the output of the ML module, while the components of our system can function separately and do not rely heavily on each others output.

Our hybrid NER system (Oudah and Shaalan 2012; Shaalan and Oudah 2014) is an integration of the rule-based approach and the ML-based approach for Arabic NER. The rule-based component is a reproduction of NERA, an Arabic rule-based NER system developed by Shaalan and Raza (2008), while the ML-based component adopts Decision Tree algorithm, which proved its efficiency as an ML algorithm in the hybrid system (Oudah and Shaalan 2012). The following sums up the distinct features of our proposed approach compared to other hybrid approaches:

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

- *Number of NE types:* Our hybrid approach differs from the Srihari *et al.* (2000) approach which handles limited types of NEs by each module or approach. In our hybrid approach, we cover a reasonable number of NE types, eleven types in particular, in a pipeline processing of rule-based and ML-based NER. The output of the rule-based component is utilized as contextual features, which is used along other language-specific and language independent features as the feature set in the ML-based component.
- *Number of modules:* In our hybrid NER, the rule-based and ML-based components work together for producing the final annotations. However, each component can function independently, in contrast with Seon *et al.* (2001)'s approach.
- *Architecture and the feature set:* Our proposed approach differs from Mencius approach (Tsai *et al.* 2004) in two aspects, including the architecture of the hybrid system and the selected feature set. Mencius is heavily dependent on the output of the rule-based module and uses it to extract the features used by the ML module, while our hybrid approach does not rely only on the rule-based system's output in extracting the features.
- *Domain dependency:* Our hybrid system is domain-independent, while Küçük and Yazici (2012)'s system is built with the capability of supporting only four specific domains, including news, financial news, historical news and child stories.
- *Role of the rule-based component:* In a comparison with Petasis *et al.* (2001) approach, our hybrid approach does not rely heavily on the rule-based decisions to extract language-independent and language-specific features of other categories, such as morphological, contextual and word-level features. Basically, the rule-based decisions are used to extract the contextual rule-based features to be added to the feature set.

Our Arabic hybrid NER system has gone through rigorous evaluation as large set of features has been explored, including rule-based features, morphological features, POS features, gazetteer features, contextual features and word-level features. According to the empirical results, our system outperforms the state-of-the-art of the Arabic NER in terms of F-measure when applied to the standard ANERcorp dataset with F-measure of 0.94 for person NEs, 0.90 for location NEs and 0.88 for organization NEs (Oudah and Shaalan 2012; Shaalan and Oudah 2014). In this study, we apply the proposed rule acquisition methodology on the output of the hybrid system in order to extract and implement new linguistic rules, which are capable of improving the performance of the rule-based component.

4 The system architecture

This study aims at improving the performance of the rule-based NER component (NERA), i.e. a component in our Hybrid NER system for Arabic. In this section, we give thorough description of the hybrid system's architecture illustrated in Figure 1.

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

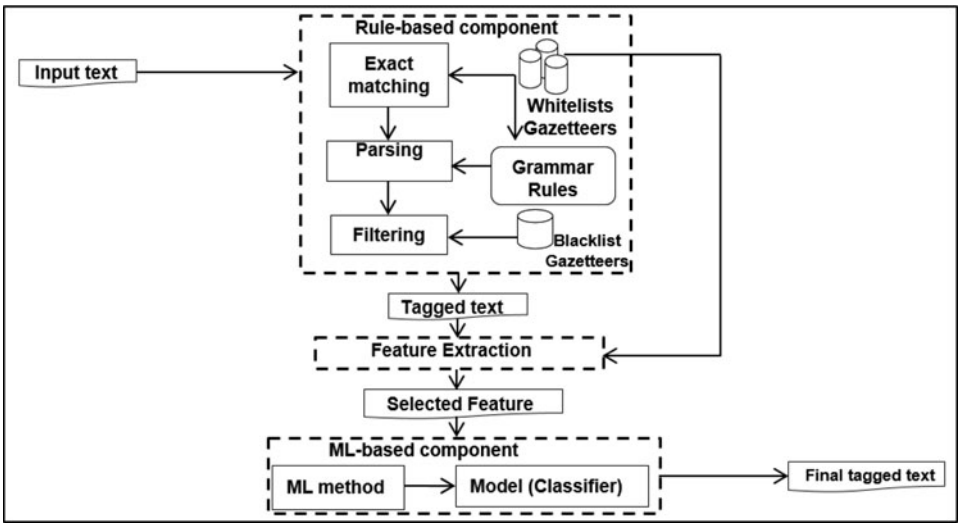


Fig. 1. The architecture of the hybrid NER system.

4.1 The rule-based component

The rule-based component is a reproduction of NERA system (Shaalan and Raza 2008), which in all aspects is domain independent, using the GATE framework[1] (Cunningham *et al.* 2011). GATE is a free open source application that allows for building applications for various NLP tasks. When it comes to NER, GATE facilitates the development of rule-based NER systems through providing the user with the capability to implement grammatical rules as finite state transducers using JAPE. JAPE is a pattern specification language, which enables the implementation of grammatical based on regular expression. The rule-based component (NERA) consists of three main modules: Whitelists (pre-defined lists of different types of NEs), Grammar Rules and a Filtration mechanism (blacklists of invalid NEs). In the GATE framework, NERA works as a corpus pipeline in which a group of documents comprising a corpus is processed through a set of processing tools, including an Arabic tokenizer, and resources, including, a set of gazetteers and local grammatical rules. NERA is built with the capability of recognizing eleven NE types, i.e. person, location, organization, date, time, measure, price, percent, ISBN, filename and phone number. In Arabic NER field, ENAMEX classic tags, i.e. NEs falling under proper names representing person, location and organization names, are commonly applied in NER systems reflecting the importance of these NEs and making the comparison of experimental results of different Arabic NER system more applicable. Therefore, in this study, we focus on improving the set of ENAMEX linguistic rules.

Moreover, a set of gazetteers is provided to the rule-based component to support the linguistic rules, which make use of gazetteers in their structures. The gazetteers are classified into categories according to the linguistic rules they support, e.g.

[1] GATE is freely available at the web link: `http://gate.ac.uk/`

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Table 1. *The number of gazetteers and rules for each NE type*

	Per.	Loc.	Org.	Date	Time	Price	Measur.	Percent	Phoneno.	Filename	ISBN	Total
Gaz.	11	20	8	12	10	8	7	3	7	3	1	90
Rules	9	20	9	7	8	4	3	1	7	3	2	73

person gazetteers. NERA contains a total of seventy-three linguistic rules and ninety gazetteers that are used by the ML-based component as well. For each set of distinct rules, e.g. person names rules, there is a set of corresponding gazetteers. Table 1 illustrates the distribution of gazetteers and rules for each NE type in our rule-based system. The system takes the input text in its original format, i.e. Modern Standard Arabic, which undergoes preprocessing and NE recognition by the rule-based component.

In this study, eleven, twenty and eight gazetteers (a total of thirty-nine) are utilized by the updated version of NERA, i.e. NERA 2.0, for person, location and organization NER, respectively, as the targeted group is ENAMEX. It is noteworthy that the set of gazetteers for each NE type also includes list of indicators/keywords for NEs, which assist the detection process via the linguistic rules.

In NERA, the condition part of each ENAMEX linguistic rule includes up to three patterns. They are able to identify:

(a) The relevant context that precedes an NE,
(b) The NE itself with its boundaries,
(c) The relevant context that follows an NE.

Where the middle pattern is mandatory. The optional patterns do not only act as indicators/clues of the NE, but also play an important role in determining the type of the NE. This will also allow for acronyms and abbreviations to be dealt with properly if they are preceded or followed by indicator keywords. The purpose of each individual rule is to label a single NE, which can span multiple tokens. It is worth noting that the NER rules are applied in the following order: person name rules, location name rules and then organization name rules. This pipeline allows rules to use the labels previously produced by rules of other NE types. We assume that a word cannot be annotated with multiple NE categories in the final output of the system. Therefore, if a word is tagged with two different NE labels another two sets of NE rules are applied to resolve this conflict such that the word is assigned the closest NE label. For instance, having a city name as part of a bank name, such as 'The National Bank of Abu Dhabi'. Our system will consider 'Abu Dhabi', i.e. a city, part of an organization NE and not a location NE.

Figure 2 illustrates a *JAPE* implementation of a linguistic rule for Person NER. This rule is composed of a simple pattern for identifying Arabic person names. It recognizes the word (possibly, a sequence of words) that is presented in the *Full_Person_Name* gazetteer, e.g. 'طه حسين' (Taha Hussein). In JAPE, Lookup.majorType refers to a look up table, i.e. a gazetteer such as Full_Person_Name, where an input word is looked up as part of the expression in brackets,

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Person Rule in the form of regular expression:

(Full Person Name)

Person Rule as implemented in GATE (i.e. in JAPE language):

```
Rule: PersonRule9
Priority:10
( {Lookup.majorType == "Full_Person_Name"} ) :Per
-->
:Per.Person= {rule="PersonRule9"}
```

Fig. 2. An Example of a rule implemented in JAPE for person NER.

Location Rule in the form of regular expression:

((ب)? + Administrative division + (Country Name | State Name | Any Location))

Location Rule as implemented in GATE:

```
Rule: LocationRule6
Priority:14
({Token.string=="ب"}{Lookup.majorType == "AdmDiv"}
({Lookup.majorType == "Country"}|{Lookup.majorType == "State"}|{Token}):Loc  )
 -->
:Loc.Location={rule="LocationRule6}
```

Fig. 3. An Example of a rule implemented in JAPE for location NER.

preceding the keyword ':Per', ':Loc' or ':Org' for identifying the resultant NE type as a person, location or organization NE, respectively. Figure 3 shows another rule for Location NER, which includes two patterns for recognizing an indicator that commonly precedes a Location name pattern. To preserve the rule space, relevant indicators of NEs are grouped together under a gazetteer where they can be looked up. Moreover, a gazetteer also groups either full or partial instances of a certain NE type. The rule presented in Figure 3 works as follows: If the expression starts with an *administrative division* indicator that might be prefixed by a clitic and is followed by either a state name, a country name or any other location listed in the location gazetteers, then the expression followed by the indicator pattern is recognized as a Location NE. Examples of Location NEs that can be recognized by this rule are (highlighted by the underlined words): مملكة البحرين (Kingdom of Bahrain) and بامبراطورية اليابان (in the Empire of Japan). It is worth noting that if the administrative division indicator is prefixed by 'الـ' (The, definite article), it is not considered an indicator but rather a part of the NE, e.g. المملكة العربية السعودية (The Kingdom of Saudi Arabia).

Moreover, Figure 4 demonstrates a third rule for recognizing organization NEs, which is comprised of three patterns: a preceding indicator pattern, an organization name pattern that could comprise multiple tokens/words and a location NE as a succeeding indicator pattern. This is an example of the previously mentioned processing order posed when applying NER task, as a location name determined by the location NER is used by the a rule of the organization NER. 'جامعة ادنبرة

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Organization Rule in the form of regular expression:

(organization preceding indicator) + (organization named entity) + (في) ? + (Location name)

Organization Rule as implemented in GATE:

```
Rule: Organization3
Priority:30
( ( {Lookup.majorType=="org_preceding_keyword"} ({Token})[1,4] ):Org
  ({Token.string=="في"})? ({Location}) )
-->
:Org.Organization = {rule="OrganizationRule3"}
```

Fig. 4. An example of a rule implemented in JAPE for organization NER.

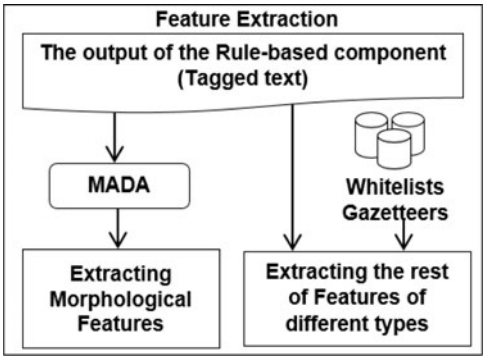


Fig. 5. The architecture of the feature extraction phase.

في المملكة المتحدة' (University of Edinburgh in the United Kingdom) is an instance of an NE extracted by this rule in which the underlined text represents the detected organization NE. The preceding indicator, i.e. University 'جامعة', is part of the organization name in the case of this rule.

4.2 The ML-based component

The ML-based component relies on two main aspects: feature engineering and ML algorithm selection. Feature engineering involves selecting a combination of classification features from the entire feature space. Figure 5 illustrates the architecture of the feature extraction phase. The features explored are divided into five categories: morphological features, POS category, gazetteer features, contextual features and word-level features. A sixth category, i.e. the rule-based features, is introduced to the feature set when the ML-based system is used as part of the hybrid approach. The morphological features are generated by MADA (Roth *et al.* 2008; Habash, Owen and Ryan 2009), which is a state-of-the-art Arabic morphological analyzer. Feature engineering allows studying the impact of different feature combinations on the overall system's performance, when several dimensions are taken into consideration, including the NE category, the ML algorithm, and the considered features. We have used WEKA (Hall *et al.* 2009), a comprehensive workbench with support for a large number of ML algorithms, as the development environment of the ML-based component. As illustrated in Figure 6, the output of

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

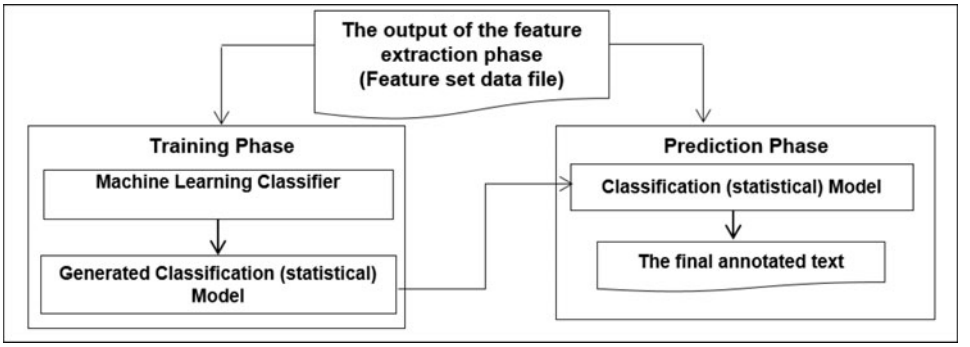


Fig. 6. The architecture of the training phase and the prediction phase.

the training phase, i.e. the classification model, is used in the prediction phase to generate the final NE annotation. The state-of-art Arabic hybrid NER (Oudah and Shaalan 2012; Shaalan and Oudah 2014) shows that the Decision Tree algorithm outperforms other algorithms, i.e. SVM and logistic regression, when adopted by the ML-based component in the context of the hybrid approach. Therefore, in this study, the output of the hybrid NER system, when Decision Tree algorithm is adopted, is analyzed and exploited to derive new patterns in order to enhance NERA.

The feature set used for ENAMEX extraction includes the following features:

- *Morphological features:* These linguistic features are derived from morphological analysis decisions, i.e. a set of thirteen features generated by MADA as explained in Table 3 (Habash, Owen and Ryan 2010). These features help distinguishing NEs from regular text based on the morphological state.
- *POS category:* The POS tag is a good distinguishing feature for Arabic NEs; it has been studied separately in the literature to determine its impact on NER. The POS feature is the POS category of the targeted word estimated by MADA (Roth *et al.* 2008; Habash *et al.* 2009) as explained in Table 3. This feature allows the classifier to learn which POS tags usually NEs occur with.
- *Nominal flag:* A binary feature to represent whether POS tag is an N (or PN). This feature supports the previous one, i.e. POS category, by stating whether a word is nominal or not.
- *Check person/location/organization gazetteers feature flags:* A binary feature to represent whether the word (or left/right neighbor of the targeted word) belongs to person/location/organization gazetteer(s). These features help in revealing the contextual pattern of ENAMEX NEs.
- *Word length flag:* A binary feature to represent whether the current word length ≥ 3. Our study indicates that the length of a word in an NE is commonly three or more characters, but in few cases it could be less than three characters e.g. person name 'مي' (Mai) (i.e. composed of two characters in Arabic). This feature allows detecting all cases.
- *Dot flag:* A binary feature to represent whether the word has an adjacent dot, e.g. beginning/end of sentence or part of an abbreviation. This feature

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Table 2. *Example of the Rule-based features for a phrase from ACE 2004 NW dataset*

Word	Transliteration*	Translation	Word N-2	Word N-1	Word N	Word N+1	Word N+2
محطات	maHaT~aAt	generating	OTHER	OTHER	OTHER	OTHER	OTHER
توليد	tawliyd	Stations	OTHER	OTHER	OTHER	OTHER	OTHER
الكهرباء	AlkahrabaA'	electricity	OTHER	OTHER	OTHER	OTHER	Location
في	fiy	in	OTHER	OTHER	OTHER	Location	OTHER
لبنان	LubnaAn	Lebanon	OTHER	OTHER	Location	OTHER	OTHER

*We used Habash–Soudi–Buckwalter transliteration scheme (Habash *et al.* 2007).

allows exploiting the word position as part of the constructed patterns in the classification model.

- *Check person/location/organization gazetteers feature flags:* A binary feature to represent whether the word (or left/right neighbor of the targeted word) belongs to person/location/organization gazetteer(s). These features help in revealing the contextual pattern of ENAMEX NEs.
- *Capitalization flag:* A binary feature to represent the existence of capitalization information on the English gloss (translation) corresponding to the analyzed Arabic word. The English gloss is additional information generated by MADA, which we utilized as a possible indication of an NE if it is capitalized.

4.3 The hybrid architecture

The integration of the rule-based and ML-based components produces the hybrid NER system, which aims at improving the overall performance of NER for Arabic. Figure 1, as previously mentioned, illustrates the hybrid architecture of our NER system. It is worth noting that the rule-based and ML-based components are independent and can operate separately (Oudah and Shaalan 2012; Shaalan and Oudah 2014).

The rule-based and ML-based component are integrated via the rule-based features. The rule-based features are the main contribution of the rule-based NER system (NERA) to the hybrid approach. They are derived from the rule-based decisions and defined in terms of a sliding window of five words centered by the candidate word. Table 2 shows examples of rule-based feature vectors. When the ML-based component is used alone, these features are completely excluded.

5 The methodology for Acquiring and applying new rules

In this section, we illustrate our methodology for acquiring new NER rules with the purpose of improving the coverage and overall performance of the rule-based NER component, i.e. NERA. Our proposed methodology for revealing new grammatical rules, which can be added to NERA as an enhancement, is guided by deep analysis of

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Table 3. *MADA morphological features with their description (Habash et al. 2010)*

No.	Feature	Feature value definition
1	Aspect	Verb aspect: command, imperfective, perfective or not applicable (NA)
2	Case	Grammatical case: nominative, accusative, genitive, NA or undefined
3	Gender	Nominal gender: feminine, masculine or NA
4	Mood	Grammatical mood: indicative, jussive, subjunctive, NA or undefined
5	Number	Grammatical number: singular, plural, dual, NA or undefined
6	Person	Person information: first, second, third or NA
7	State	Grammatical state: indefinite, definite, construct/poss/Idafa, NA or undefined
8	Voice	Verb voice: active, passive, NA or undefined
9	Proclitic 3	Question proclitic: no proclitic (NP), NA or interrogative particle *¿ a*
10	Proclitic 2	Conjunction proclitic: NP, NA, conjunction *fa*, response conditional *fa*, subordinating conjunction *fa*, conjunction *wa*, particle *wa* or subordinating conjunction *wa*
11	Proclitic 1	Preposition proclitic: NP, NA, particle *bi*, preposition *bi*, preposition *ka*, emphatic particle *la*, preposition *la*, response conditional *la*, jussive *li*, preposition *li*, future marker *sa*, preposition *ta*, particle *wa*, preposition *wa*, preposition *fy*, negative particle *lA*, negative particle *mA*, vocative *yA*, vocative *wA* or vocative *hA*
12	Proclitic 0	Article proclitic: NP, NA, determiner, negative particle *lA*, negative particle *mA*, relative pronoun *mA* or particle *mA*
13	Enclitics	Pronominal: No enclitic, NA, first person (*plural*|*singular*), second person (dual—(feminine (*plural*|*singular*))—(masculine (*plural*|*singular*))), third person (dual—(feminine (plural —singular))—(masculine (*plural*|*singular*))), vocative particle, negative particle *lA*, interrogative pronoun (*ma*|*mA*|*man*), relative pronoun (*ma*|*mA*|*man*) or subordinating conjunction (*ma*|*mA*)
	POS	POS definition: nouns, number words, proper nouns, adjectives, adverbs, pronouns, verbs, particles, prepositions, abbreviations, punctuation, conjunctions, interjections, digital numbers or foreign/latin

the hybrid NER system's output of annotations. The rule acquisition methodology consists of three main steps:

(1) Extracting the feature vectors of NEs misclassified by NERA, but correctly classified by the hybrid system.
(2) Analyzing the output of the previous step to reveal candidate patterns for NER.
(3) Assessing the performance of candidate patterns to find the most promising ones.

In this study, we apply the proposed methodology on ACE 2004 NW[2] dataset (Mitchell *et al.* 2005). The ACE 2004 NW dataset is a standard Arabic dataset that was created with NE tags by LDC to be utilized for NER evaluation. The

[2] Available for us under license agreement from the LDC.

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Table 4. *A sample of NE tags correctly classified by the hybrid NER system but misclassified by the pure rule-based (RB) NER component when applied on ACE 2004 NW dataset — the candidate word is underlined*

Sentence in Arabic	Translation	RB label	Hybrid label
زيارة عمل قصيرة للاردن	Short work visit to Jordan	Other	Location
رئيس الوزراء الكندي جان كريتيان	Canadian Prime Minister Jean Chretien	Other	Person
التوصل الى سلام عادل	To reach a justice peace	Person	Other
لقاء ابناء المدينة الواحدة	The meeting of the sons of the one city	Location	Other
الرئيس الكيني دانيال اراب موي	Kenyan President Daniel Arap Moi	Other	Person

misclassifications made by NERA are calculated via GATE, which has its own built-in evaluation tool that makes use of the reference dataset, i.e. ACE 2004 NW, in the evaluation process. The NERA performance is evaluated in terms of precision, recall and F-measure. Whereas, the performance of the hybrid system is evaluated via WEKA workbench, in terms of precision, recall and F-measure as well. Both GATE and WEKA's evaluation modules produce indexed lists of correctly classified and misclassified words; hence, the process of finding the set of intersection between the correctly classified NEs by the hybrid system and the misclassified ones by NERA is very simple. Therefore, the whole process of evaluation is fully automated. Table 4 shows examples of NEs that are misclassified by NERA, but correctly classified by the hybrid system. The performance of the first version of the rule-based system (NERA) and the Hybrid system, when applied on ACE 2004 NW, is illustrated in Table 5. Ten-fold cross validation is used for training and evaluation in order to avoid overfitting. The overall false positive and false negative rates achieved by NERA, when applied to ACE 2004 NW, are 0.042 and 0.381, respectively. As shown in Table 5, the hybrid system outperforms NERA and the ML-based component, separately, in terms of precision and recall and F-measure. The number of NEs misclassified by NERA, but correctly classified by the hybrid system is 2,161 out of 2,448. By applying the proposed rule acquisition methodology on NERA's misclassifications, the POS category and gazetteers features are found to be of significance for characterizing the misclassified NEs and hence, for building the enhanced linguistic rules as well. In the literature, good NER performance is reported when the POS tags are considered (Oudah and Shaalan 2012; Oudah and Shaalan 2013). Moreover, the shared task of CoNLL now provides a POS column in its corpora to serve the same purpose. The second step in our rule acquisition methodology reveals that 72.14 per cent and 16.33 per cent of the misclassified NEs are PN and N, respectively, as illustrated in Figure 7.

Abdallah *et al.* (2012) has reported that a typical length of an NE of ENAMEX type is three words long. Therefore, we use a window of three parts, i.e. immediate left neighbor, candidate NE and immediate right neighbor, to study the gazetteer

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Table 5. *A sample of NE tags correctly classified by the hybrid NER system but misclassified by the pure rule-based (RB) NER component when applied on ACE 2004 NW dataset (the candidate word is underlined)*

System	Per.			Loc.			Org.		
	P	R	F	P	R	F	P	R	F
NERA (rule-based component)	0.34	0.35	0.35	0.63	0.72	0.67	0.43	0.35	0.38
ML-based component	0.79	0.84	0.81	0.79	0.81	0.80	0.74	0.48	0.58
Hybrid system	0.83	0.82	0.82	0.84	0.84	0.84	0.82	0.49	0.62

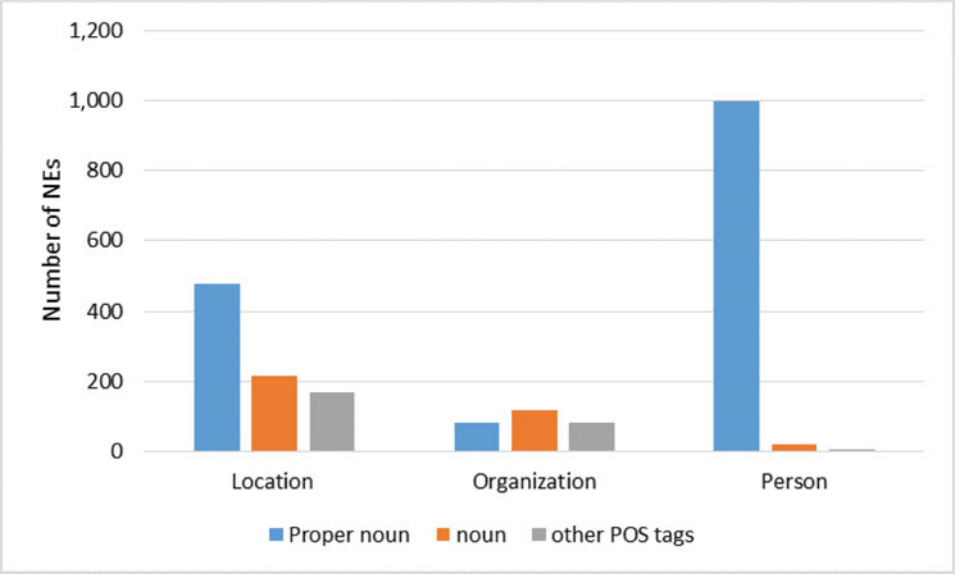


Fig. 7. (Colour online) The distribution of the POS tags of the person, location and organization NEs (i.e. noun and proper noun) in ACE 2004 NW dataset.

property in misclassified NEs. The results show that 43.14 per cent of the candidate NEs belong to gazetteers of their category, 33.17 per cent of the immediate left neighbors and 23.23 per cent of the immediate right neighbors belong to the same gazetteer category as the candidate NE, as illustrated in Figure 8.

As a result of this feature analysis, we propose a new form of linguistic patterns, which incorporates the following distinctive features:

- *POS tag:* represents whether a candidate word is N or PN.
- *Gazetteer of candidate word:* represents the gazetteer category (person, location or organization) a candidate word belongs to (if any).
- *Left-neighbor gazetteer:* represents the gazetteer category the immediate left neighbor belongs to (if any).
- *Right-neighbor gazetteer:* represents the gazetteer category the immediate right neighbor belongs to (if any).

The proposed form of linguistic rules covers the following cases:

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

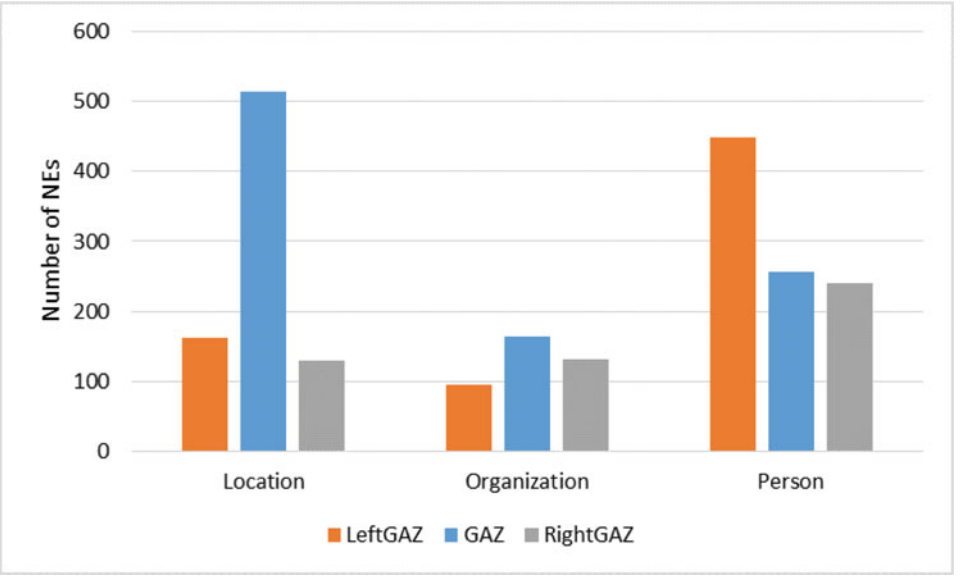


Fig. 8. (Colour online) The distribution of the gazetteer features of the person, location and organization NEs (i.e. of the candidate word and its immediate left and right neighbors) in ACE 2004 NW dataset.

- If a candidate word is an N or PN, belongs to a gazetteer of a certain category, and at least one of its neighbors belongs to a gazetteer of the same gazetteer category as the candidate word, then the candidate word will be assigned that NE category as well as the neighbor(s) with same gazetteer category.
- If a candidate word is N or PN, but does not belong to any gazetteer, then the neighbors' gazetteer category, if any, will determine the NE label assigned to that candidate word as follows:
  - — If both neighbors belong to gazetteers of the same category, then the candidate word and the neighbors are assigned that NE category.
  - — If only one of the neighbors belongs to a gazetteer of a certain category, then the candidate word and that neighbor are assigned that NE category.
  - — If the neighbors belong to gazetteers of different categories, then the candidate is not identified as an NE.

Fourteen new grammatical rules can be formulated via the new form of linguistic patterns for person, location and organization NER.

They are broadly classified under four groups based on the candidate word characteristics, including its POS tag, i.e. whether it is N or PN, and its gazetteer membership, i.e. whether it belongs to a gazetteer in the system or not, as illustrated by hexagon shapes in Figure 9. As previously mentioned, the new form of linguistic patterns used a window of three words: the left neighbor gazetteer feature, the candidate word POS and gazetteer features and the right neighbor gazetteer feature, respectively. The fourteen new rule patterns are demonstrated in Table 6; where: LeftGaz (+/−) refers to whether or not the gazetteer category of the left neighbor is the same as the gazetteer category of the candidate word (or the right neighbor);

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

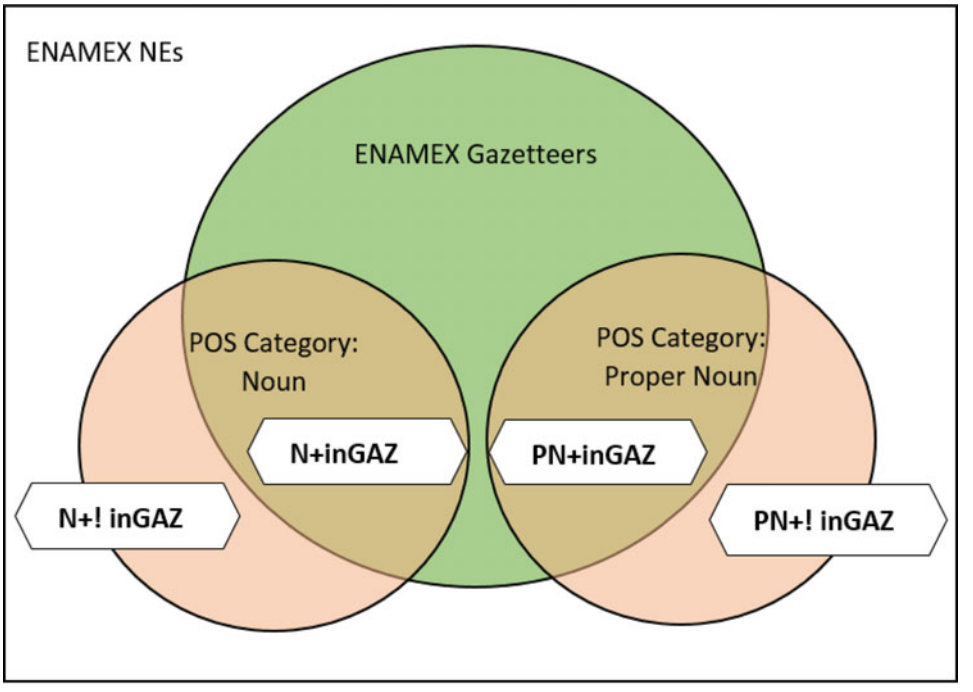


Fig. 9. (Colour online) The four groups of which the new constructed linguistic rules are distributed among based on the candidate word characteristics, i.e. its POS tag (noun (N) or proper noun (PN)), and whether it belongs to a gazetteer or not.

CandidateWord (PN/N, +/−) refers to whether the candidate word is either a PN or an N and whether or not it belongs to a gazetteer of a certain category; RightGaz (+/−) refers to whether or not the gazetteer category of the right neighbor is the same as category as the gazetteer category of the candidate word (or the left neighbor). Figure 10 demonstrates one of the new ENAMEX linguistic rules for person name recognition, which follows the new rule pattern number 9. The rule states that if a candidate word with POS tag of 'PN' does not belong to any gazetteer and its immediate right and left neighbours are presented in person name gazetteers, then the rule-based system will recognize the word and its immediate neighbours as person NE. An example of an NE detected by this rule is محمد حماد ياسين (Mohamed Hammaad Yaseen), where 'حماد' is a person name that does not belong to any of the person gazetteers and its POS category is PN.

6 Experiments

In this section, we present the experimental analysis of the performance of the rule-based component after implementing the new rules, i.e. NERA 2.0.

6.1 Experimental settings

A number of experiments have been conducted to evaluate the performance of the updated rule-based component (NERA 2.0), which has the fourteen new rule

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Table 6. *The fourteen new rule patterns in NERA 2.0*

Linguistic pattern	LeftGaz (+/−)	CandidateWord (PN/N, +/−)	RightGaz (+/−)
PN+inGAZ (1)	+	PN, +	+
PN+inGAZ (2)	+	PN, +	−
PN+inGAZ (3)	−	PN, +	+
PN+inGAZ (4)	−	PN, +	−
N+inGAZ (5)	+	N, +	+
N+inGAZ (6)	+	N, +	−
N+inGAZ (7)	−	N, +	+
N+inGAZ (8)	−	N, +	−
PN+!inGAZ (9)	+	PN, −	+
PN+!inGAZ (10)	+	PN, −	−
PN+!inGAZ (11)	−	PN, −	+
N+!inGAZ (12)	+	N, −	+
N+!inGAZ (13)	+	N, −	−
N+!inGAZ (14)	−	N, −	+

New ENAMEX Rule (i.e. for Person NE) in the form of regular expression:

((First name | Middle name | Last name) + (POS tag="proper noun") + (First name | Middle name | Last name))

New ENAMEX Rule (i.e. for Person NE) as implemented in GATE:

```
Rule: PersonRule09
Priority:14
(({Lookup.majorType == "Firsts"}|{Lookup.majorType == "Middle"}|{Lookup.majorType == "Lasts"})
({Token.category == "proper_noun", Lookup.majorType != "Firsts", Lookup.majorType != "Middle",
Lookup.majorType != "Lasts"})
({Lookup.majorType == "Firsts"}|{Lookup.majorType == "Middle"}|{Lookup.majorType == "Lasts"}))
:Per
-->
:Per.Person={rule="PersonRule09}
```

Fig. 10. Example of a new person NE rule (i.e. based on new rule #09).

patterns implemented for person, location and organization NER. The dataset used for evaluation is a POS-tagged version of ACE 2004 NW dataset, i.e. estimated by MADA. It contains a total of 6,627 ENAMEX NEs: 1,865 person names, 3,449 location names and 1,313 organization names. The gazetteers utilized by NERA 2.0 contain 4,297 person names, 1,204 location names and 1,257 organization name. And as previously mentioned, the number of NEs misclassified by NERA, when applied to ACE 2004 NW, but correctly classified by the hybrid system is 2,161 (1,021 for person, 860 for location and 280 for organization) out of 2,448 NERA misclassifications.

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Table 7. *The results of applying NERA 2.0 on ACE 2004 NW dataset in terms of improvement. The numbers indicate the ratio of correctly classified NEs by NERA 2.0 over the misclassified ones by NERA*

Linguistic pattern	Person	Location	Organization
PN+inGAZ (1)	3.13 per cent	0.81 per cent	0 per cent
PN+inGAZ (2)	5.48 per cent	6.04 per cent	0.71 per cent
PN+inGAZ (3)	4.40 per cent	1.97 per cent	0.71 per cent
PN+inGAZ (4)	12.04 per cent	33.02 per cent	3.92 per cent
N+inGAZ (5)	0 per cent	1.74 per cent	8.21 per cent
N+inGAZ (6)	0 per cent	0 per cent	3.21 per cent
N+inGAZ (7)	0 per cent	7.90 per cent	23.57 per cent
N+inGAZ (8)	0 per cent	0.11 per cent	4.64 per cent
PN+!inGAZ (9)	6.26 per cent	0.58 per cent	0.71 per cent
PN+!inGAZ (10)	28.40 per cent	0.23 per cent	6.42 per cent
PN+!inGAZ (11)	9.40 per cent	0.58 per cent	1.78 per cent
N+!inGAZ (12)	0.19 per cent	0 per cent	0 per cent
N+!inGAZ (13)	0.48 per cent	0.34 per cent	0.35 per cent
N+!inGAZ (14)	0.19 per cent	0.23 per cent	0 per cent
Total	69.93 per cent	57.09 per cent	54.28 per cent

6.2 Results

The performance of NERA 2.0, when applied to ACE 2004 NW dataset, is presented with respect to the fourteen new rule patterns, which are to be then ranked accordingly. The empirical results show that each group of patterns has a different degree of impact on the performance, as shown in Table 7. Through integrating the new patterns, the rule-based component becomes capable of correctly recognizing 69.93 per cent, 57.09per cent and 54.28 per cent of person, location and organization NEs, respectively, out of NERA's misclassifications.

For person NER, the highest contribution to the overall performance improvement (69.93 per cent) is achieved via PN+!inGAZ group with sixty-three per cent of the correctly recognized names out of NERA's misclassified person names. The next contribution to the overall performance improvement is made by the PN+inGAZ group with 35.80 per cent, and finally N+!inGAZ group with 1.25 per cent. The results show that the majority of these person NEs are not present in our system's person gazetteers, which may contribute to them being previously misclassified by NERA, and are PN.

As for Location NER, the highest contribution to the overall performance improvement (57.09 per cent) is achieved via PN+inGAZ group with 73.31 per cent of the correctly recognized names out of NERA's misclassified location names. N+inGAZ, PN+!inGAZ and N+!inGAZ groups have contributed with 17.10 per cent, 8.55 per cent and 1.01 per cent, respectively. According the results, the majority of correctly classified location NEs are PN as well. When it comes to organization NER, the contributions to the overall performance improvement (54.28 per cent), in

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

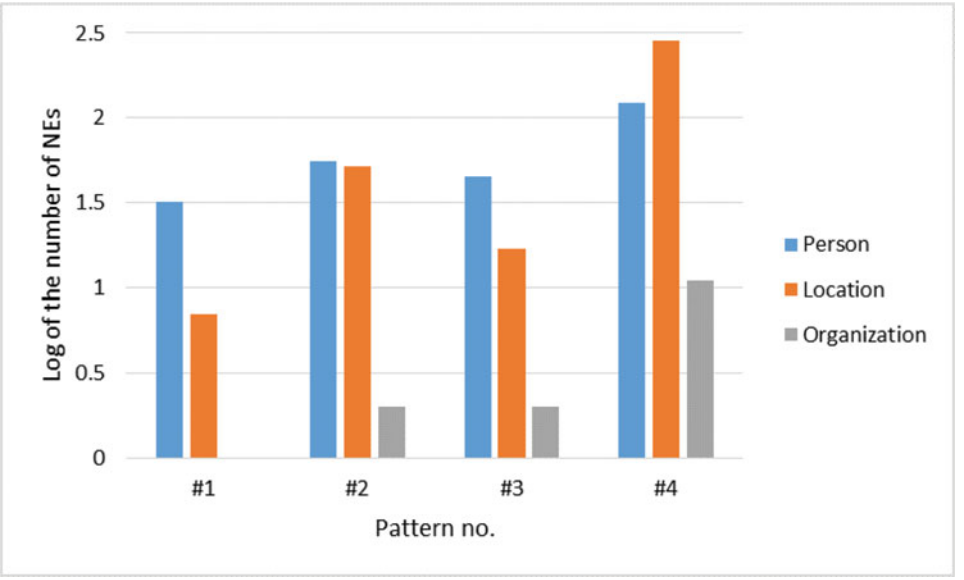


Fig. 11. (Colour online) The results of applying the PN+inGAZ group patterns on ACE 2004 NW dataset in terms of the log−10 of correctly classified NEs out of NERA's misclassifications.

descending order, are 73.10 per cent, 16.42 per cent, 9.84 per cent and 0.64 per cent of the correctly recognized names out of NERA's misclassified organization names by N+inGAZ, PN+!inGAZ, PN+inGAZ and N+!inGAZ groups, respectively. It can be noted that the majority of correctly classified organization names are with N POS tag, in contrast with the other two NE types.

As a conclusion, incorporating the POS feature of the candidate word, in addition to the neighbors' gazetteer features, in the structure of the new ENAMEX rules improves the performance, dramatically. According to Figure 11, rule pattern #4, as described in Section 5, outperforms the other patterns in the PN+inGAZ group achieving the highest performance in location, person and organization NER, respectively. This rule pattern helps overcoming the problem of having out-of-date gazetteers via examining the POS feature of the candidate word and whether it belongs to the system's gazetteers. The descending order of the rule patterns in PN+inGAZ group according to their performance is #4, #2, #3 and #1, respectively.

The N+inGAZ group's rule patterns fail to improve person NER at all, which indicates that no person name with an N POS and present in the person gazetteers is encountered in the experiments. However, these patterns are capable of improving the location and organization NER, as shown in Figure 12. The descending order of rule patterns in N+inGAZ group according to their performance is #7, #5, #8 and #6, respectively. The rule patterns in PN+!inGAZ group show promising impact on person NER with 87.04 per cent of correct person NE classifications made only by this group, as demonstrated in Figure 13. Based on further analysis, the presence of NEs with spelling variants and foreigner (transliterated) names in the Arabic text, which do not belong to the gazetteers, pose the additional challenge to NER.

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

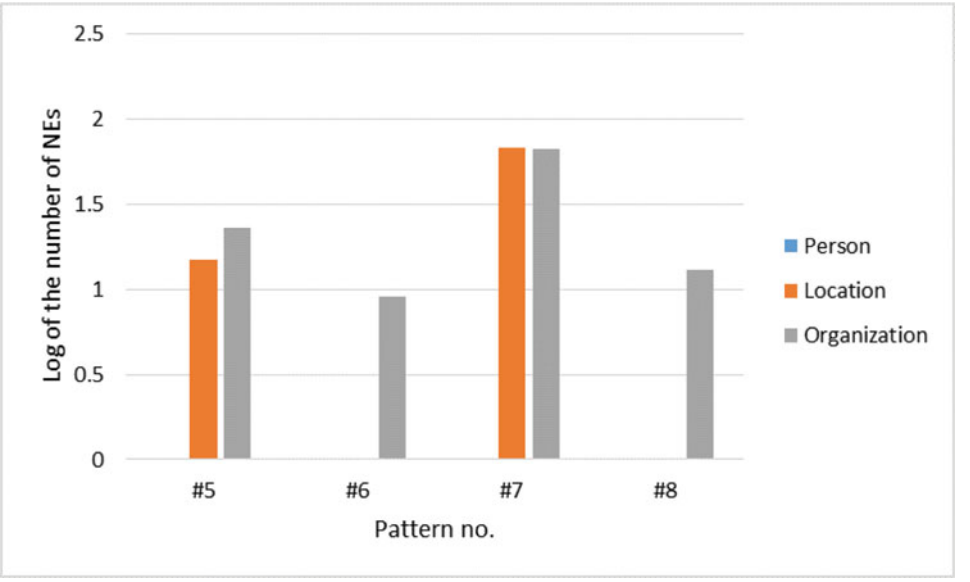


Fig. 12. (Colour online) The results of applying the N+inGAZ group patterns on ACE 2004 NW dataset in terms of the log−10 of correctly classified NEs out of NERA's misclassifications.

Therefore, we try to tackle this problem by considering the POS tag of the words and the surrounding context in the rules. The descending order of rule patterns in PN+!inGAZ group according to their performance is #10, #11 and #9, respectively.

As illustrated in Figure 14, N+!inGAZ group of rule patterns achieves a slight improvement in the performance. This observation indicates that if a candidate word is an N and does not belong to any gazetteer, then it is unlikely an NE. The descending order of patterns in N+!inGAZ group according to their performance is #13, #14 and #12, respectively.

Taking into consideration the performance of each group of new rule patterns, the rule patterns with the highest performance are patterns #4, #7, #10 and #13, respectively. As previously mentioned, incorporating the POS feature of the candidate word and the neighbors' gazetteer features, in the structure of the new ENAMEX rules helps improving the NER performance, which results in NERA 2.0 outperforming NERA. Moreover, the overall false positive and false negative rates achieved of NERA 2.0 when applied to ACE 2004 NW dataset are 0.049 and 0.151, respectively, which is an improvement when compared to the NERA where the false positive and false negative rates are 0.042 and 0.381, respectively.

7 Conclusion

In this article, we introduce a novel methodology for acquiring new linguistic NER rules in order to improve the performance of rule-based component in our Hybrid NER system for Arabic. An automated mechanism for rule-based system's maintenance will help eliminating the main disadvantage of the rule-based

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

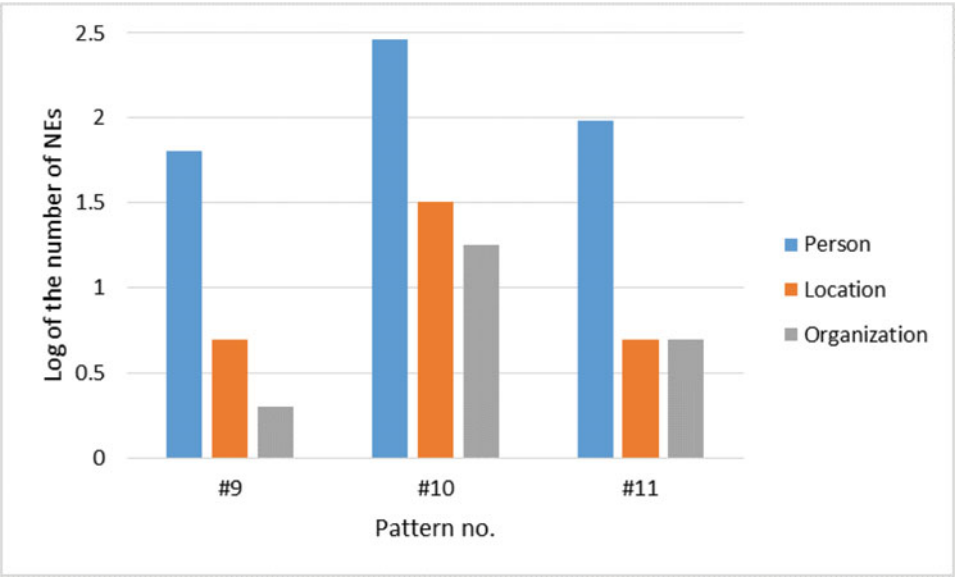


Fig. 13. (Colour online) The results of applying the PN+!inGAZ group patterns on ACE 2004 NW dataset in terms of the log−10 of correctly classified NEs out of NERA's misclassifications.

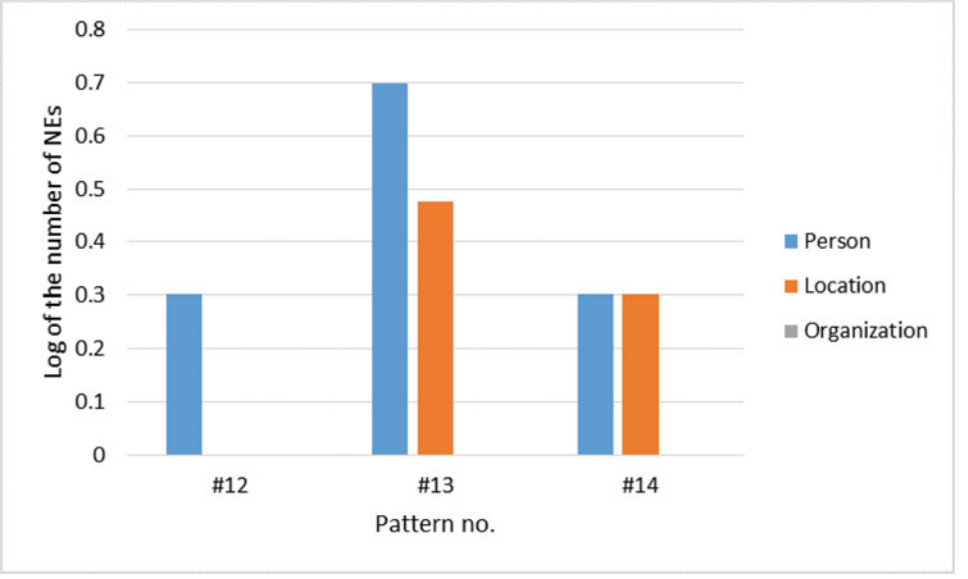


Fig. 14. (Colour online) The results of applying the N+!inGAZ group patterns on ACE 2004 NW dataset in terms of the log−10 of correctly classified NEs out of NERA's misclassifications.

approach, i.e. being labor-intensive and time-consuming when it comes to updating the system. With an improved performance of the rule-based component, more reliable and accurate results can be achieved. Our hybrid NER system is composed of two components: rule-based component (NERA) and ML-based component. It

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

is capable of recognizing eleven different types of NEs, including person, location, organization, date, time, price, measurement, per cent, phone number, ISBN and filename. The category of NEs of interest in this study is ENAMEX, i.e. person, location and organization names.

The proposed methodology consists of three main steps: (1) Extracting the feature vectors of NEs misclassified by NERA, but correctly classified by the hybrid system, (2) analyzing the output of the previous step to reveal candidate patterns for NER and (3) assessing the performance of candidate patterns to find the most promising ones. ACE 2004 NW dataset is used as a linguistic resource dataset for extracting and deriving new linguistic rules for person, location and organization names recognition due to its significance as one of the largest standard dataset used by Arabic NER research community. As a result, fourteen new patterns are produced for ENAMEX. The new patterns are distributed among four categories taking into account two factors: (1) the POS tag of the candidate word, i.e. a PN or a common N, and (2) its presence status in the system's gazetteers. We conduct a number of experiments to evaluate the performance of NERA after implementing the new patterns, i.e. NERA 2.0. The testing dataset used for evaluation is a POS tagged version of the ACE 2004 NW dataset. According to the experimental results, NERA 2.0 is capable of obtaining a coverage improvement of 69.93 per cent, 57.09 per cent and 54.28 per cent for person, location and organization names, respectively. NERA 2.0 recognizes a wide range of NERA misclassifications resulting in a decreased false negative rate with a slightly increased false positive rate that is due to the common words (NEs/keywords) across the gazetteers of different categories. This emphasizes on the need to improve the gazetteers and keep them up-to-date. According to the experimental analysis, incorporating the POS feature of the candidate word and the neighbors' gazetteer features in the structure of the new ENAMEX rules helps improving the NER performance and results in NERA 2.0 outperforming NERA.

For further research, we intend to improve the quality of the gazetteers, and explore more ML algorithms and study their impact on the overall performance of the hybrid NER system. Furthermore, we intend to apply and test the proposed methodology on other types of datasets, such as Prophet Mohammad's Hadith.

References

Abdallah, S., Shaalan, K., and Shoaib, M. 2012. Integrating rule-based system with classification for arabic named entity recognition. In *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Springer-Verlag, Berlin Heidelberg, pp. 311–22.

AbdelRahman, S., Elarnaoty, M., Magdy, M., and Fahmy A. 2010. Integrated machine learning techniques for arabic named entity recognition. *International Journal of Computer Science Issues (IJCSI)* **7**(3): 27–36.

Abdul-Hamid, A., and Darwish, K. 2010. Simplified feature set for arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop (ACL 2010)*, Stroudsburg, PA, USA, pp. 110–5.

Aboaoga, M., and Aziz, M. J. A. 2013. Arabic person names recognition by using a rule based approach. *Journal of Computer Science* **9**(7): 922–7.

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Abouenour, L., Bouzoubaa, K., and Rosso, P. 2012. IDRAAQ: new arabic question answering system based on query expansion and passage retrieval. *CLEF (Online Working Notes/Labs/Workshop)*.

Alias-I. 2008. 'LingPipe 4.1.0'. http://alias-i.com/lingpipe (accessed October 2012).

Al-Sughaiyer, I., and Al-Kharashi, A. 2004. Arabic morphological analysis techniques: a comprehensive survey. *Journal of the American Society for Information Science and Technology* **55**(3): 189–213.

Azmi, A. M., and Badia, N. 2010. e—Narrator — An application for creating an ontology of Hadiths narration tree semantically and graphically. *The Arabian Journal for Science and Engineering* **35**(2C): 51–68.

Babych, B., and Hartley, A. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT (EAMT 2003)*, Stroudsburg, PA, USA, pp. 1–8.

Benajiba, Y., Diab, M., and Rosso, P. 2008a. Arabic named entity recognition using optimized feature sets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, Stroudsburg, PA, USA, pp. 284–93.

Benajiba, Y., Diab, M., and Rosso, P. 2008b. Arabic named entity recognition: an svm-based approach. In *Proceedings of Arab International Conference on Information Technology (ACIT 2008)*, Hamamamet, Tunisia, pp. 16–8.

Benajiba, Y., Diab, M., and Rosso, P. 2009a. Arabic named entity recognition: a feature-driven study. *IEEE Transactions on Audio, Speech, and Language Processing* **17**(5): 926–34.

Benajiba, Y., Diab, M., and Rosso, P. 2009b. Using language independent and language specific features to enhance arabic named entity recognition. *The International Arab Journal of Information Technology* **6**(5): 464–73.

Benajiba, Y., and Rosso, P. 2007. ANERsys 2.0: conquering the ner task for the arabic language by combining the maximum entropy with POS-tag information. In *Proceedings of Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence (IICAI-2007)*, Pune, India, pp. 1814–23.

Benajiba, Y., and Rosso, P. 2008. Arabic named entity recognition using conditional random fields. In *Proceedings of Workshop on HLT & NLP within the Arabic World (LREC 2008)*, Marrakech, Morocco, pp. 143–53.

Benajiba, Y., Rosso, P., and Bened'I, J. M. 2007. ANERsys: an arabic named entity recognition system based on maximum entropy. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2007)*, Springer-Verlag, Berlin, Heidelberg, pp. 143–53.

Collins, M. 2002. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, pp. 1–8.

Cowie, J., and Wilks, Y. 1996. Information extraction. *Communications of the ACM* **39**(1): 80–91.

Cunningham, H., et al. 2011. *Text Processing with GATE (Version 6)*. University of Sheffield Department of Computer Science, Gateway Press CA, USA.

Darwish, K., and Magdy, W. 2014. Arabic information retrieval. *Foundations and Trends in Information Retrieval* **7**(4): 239–342.

Elsebai, A., Meziane, F., and BelKredim, F. Z. 2009. A rule based persons names arabic extraction system. *Communications of the IBIMA* **11**(6): 53–9.

Farber, B., Freitag, D., Habash, N., and Rambow, O. 2008. Improving NER in arabic using a morphological tagger. In *Proceedings of Workshop on HLT & NLP within the Arabic World (LREC 2008)*, Marrakech, Morocco, pp. 2509–14.

Farghaly, A., and Shaalan, K. 2009. Arabic natural language processing: challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)* **8**(4): 1–22.

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Finkel, J., and Manning, C. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, pp. 141–50.

Habash, N., Owen, R., and Ryan, R. 2009. MADA+TOKAN: a toolkit for arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt, pp. 102–10.

Habash, N., Owen, R., and Ryan, R. 2010. MADA+TOKAN Manual. Technical Report CCLS-10-01, Center for Computational Learning Systems (CCLS), Columbia University.

Habash, N., and Rambow, O. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Stroudsburg, PA, USA, pp. 573–80.

Habash, N., Soudi, A., and Buckwalter, T. 2007. On arabic transliteration. *Arabic Computational Morphology: Knowledge-based and Empirical Methods* **38**: 15–22.

Habash, N. Y. 2010. *Introduction to Arabic Natural Language Processing*. Mogran & Claypool Publisher, San Rafael, California, USA.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* **11**(1): 10–8.

Hamadene, A., Shaheen, M., and Badawy, O. 2011. ARQA: an intelligent arabic question answering system. In *Proceedings of Arabic Language Technology International Conference (ALTIC 2011)*, Alexandria, Egypt, pp. 1–9.

Küçük, D., and Yazici, A. 2012. A hybrid named entity recognizer for Turkish. *Expert Systems with Applications* **39**(3): 2733–42.

Maloney, J., and Niv, M. 1998. TAGARAB: a fast, accurate arabic name recognizer using high-precision morphological analysis. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (Semitic 1998)*, Stroudsburg, PA, USA, pp. 8–15.

Mayfield, J., McNamee, P., and Piatko, C. 2003. Named entity recognition using hundreds of thousands of features. In *Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003 (CONLL 2003)*, Stroudsburg, PA, USA, pp. 184–7.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., and Wilks, Y. 2001. Named entity recognition from diverse text types. In *Proceedings of Recent Advances in Natural Language Processing 2001 Conference*, Tzigov Chark, Bulgaria.

Mesfar, S. 2007. Named entity recognition for arabic using syntactic grammars. In *Proceedings of the 12th International Conference on Application of Natural Language to Information Systems, Springer-Verlag*, Berlin, Heidelberg, pp. 305–16.

Mitchell, A., Strassel, S., Huang, S., and Zakhary, R. 2005. *ACE 2004 Multilingual Training Corpus*, Philadelphia, PA: Linguistic Data Consortium.

Mohammed, N. F., and Omar, N. 2012. Arabic named entity recognition using artificial neural network. *Journal of Computer Science* **8**(8): 1285–93.

Nadeau, D., and Sekine, S. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* **30**(1): 3–26.

Oudah, M., and Shaalan, K. 2013. *Person Name Recognition Using the Hybrid Approach,* vol. 7934, *Lecture Notes in Computer Science, Natural Language Processing and Information Systems*. Berlin Heidelberg: Springer, pp. 237–48.

Oudah, M. M., and Shaalan, K. 2012. A pipeline arabic named entity recognition using a hybrid approach. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, pp. 2159–76.

Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., and Spyropoulos, C. D. 2001. Using machine learning to maintain rule-based named-entity recognition and classification systems. In *Proceeding Conference of Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 426–33.

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press

Riaz, K. 2010. Rule-based named entity recognition in urdu. In *Proceedings of the 2010 Named Entities Workshop (ACL 2010)*, Stroudsburg, PA, USA, pp. 126–35.

Roth, R., Rambow, O., Habash, N., Diab, M., and Rudin, C. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of ACL-08: HLT, Short Papers*, Stroudsberg, PA, USA, pp. 117–20.

Salloum, W., and Habash, N. 2012. Elissa: a dialectal to standard arabic machine translation system. In *Proceedings of the International Conference on Computational Linguistics*, Mumbai, India, pp. 385–92.

Seon, C., Ko, Y., Kim, J., and Seo, J. 2001. Named entity recognition using machine learning methods and pattern-selection rules. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, pp. 229–36.

Shaalan, K. 2010. Rule-based approach in arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)* **3**(3): 11–9.

Shaalan, K. 2014. A survey of arabic named entity recognition and classification. *Computational Linguistics* **40**(2): 469–80.

Shaalan, K., and Oudah, M. 2014. A hybrid approach to arabic named entity recognition, *Journal of Information Science (JIS)* **40**(1): 67–87.

Shaalan, K., and Raza, H. 2007. Person name entity recognition for arabic. In *Proceedings of the 5th Workshop on Important Unresolved Matters*, Prague, Czech Republic, pp. 17–24.

Shaalan, K., and Raza, H. 2008. Arabic named entity recognition from diverse text types. In *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL 2008)*, Berlin, Heidel-berg: Springer-Verlag, pp. 440–51.

Shaalan, K., and Raza, H. 2009. NERA: named entity recognition for arabic. *Journal of the American Society for Information Science and Technology* **60**(8): 1652–63.

Shaalan, K., Abo Bakr, H., and Ziedan, I. 2009. A hybrid approach for building arabic diacritizer. In *Proceedings of the 12th European Chapter of the Association for Computational Linguistics (EACL 2009) Workshop on Computational Approaches to Semitic Languages, Association for Computational Linguistics*, Athens, Greece, pp. 27–35.

Shaalan, K., Monem, A., and Rafea, A. 2007. Arabic morphological generation from interlingua: a rule-based approach. In *Proceedings of the 4th International Conference on Intelligent Information Processing*, Adelaide, Australia, pp. 441–51.

Srihari, R., Niu, C., and Li, W. 2000. A hybrid approach for named entity and sub-type tagging. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLC 2000)*, Stroudsburg, PA, USA, pp. 247–54.

Tsai, T., Wu, S., Lee, C., Shih, C., and Hsu, W. 2004. Mencius: a Chinese named entity recognizer using the maximum entropy-based hybrid model. *Computational Linguistics and Chinese Language Processing* **9**(1): 65–82.

Zaghouani, W. 2012. RENAR: a rule-based arabic named entity recognition system. *ACM Transactions on Asian Language Information Processing* **11**(1): 1–13.

Zhou, G., and Su, J. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Stroudsburg, PA, USA, pp. 473–80.

Zirikly, A., and Diab, M. 2015. Named entity recognition for arabic social media. In *Proceedings of NAACL-HLT 2015*, Denver, Colorado, USA, pp. 176–85.

https://doi.org/10.1017/S1351324916000097 Published online by Cambridge University Press